

ASSIGNING POLYMORPHIC ENDOGENOUS RETROVIRUS INTEGRATION SITES USING A MIXTURE MODEL

BY LE BAO¹ AND DAVID R. HUNTER¹ AND MARY POSS²

*Department of Statistics¹, Penn State University, University Park, PA,
USA*

Department of Biology², Penn State University, University Park, PA, USA

E-mail: lebao@psu.edu; dhunter@stat.psu.edu; mposs@bx.psu.edu

Pennsylvania State University

Structural variation occurs in the genomes of individuals because of the different positions occupied by repetitive genome elements such as endogenous retroviruses, or ERVs. The presence or absence of ERVs can be determined by identifying the junction with the host genome using high-throughput sequence technology. The resulting data are a matrix giving the number of sequence reads assigned to each ERV-host junction sequence for each sampled individual. We present a novel two-component mixture of negative binomial distributions to model these counts and to assign a probability that a given ERV is present in a given individual, giving statistical and biological rationale for our choice. We explain ways in which our approach is superior to existing alternatives, including another form of two-component mixture model and the much more common approach of selecting a threshold count for declaring the presence of an ERV. We apply our method to a data set of ERV integrations in mule deer [*Odocoileus hemionus*] from Oregon, Montana and Wyoming and blacktail deer from Oregon. In addition, we use the data to determine the relatedness of mule deer and find that mule deer from Oregon and Montana are more closely related to each other than they are to Wyoming animals based on the profile of shared virus integration sites.

1. Introduction. Determining how genome sequences vary among individuals and populations is an important research area because genetic differences can confer phenotypic differences. The most commonly reported variations in genome sequence between two individuals are those that occur at the nucleotide level, e.g, single nucleotide polymorphisms (SNPs). These are typically identified by comparing the nucleotide at each position of a query sequence to that of a reference genome. Individual genomes can also differ in the relative position and number of homologous genome regions.

Keywords and phrases: EM algorithms; Negative Binomial; Read Count Data

For example, a genetic locus can be duplicated, deleted, inverted or moved to a new location in one genome compared to another. These changes in the genome are called genome structural variations (GSVs) and are more difficult to analyze than SNPs, particularly if a region is present in the query but absent from the reference. Transposable elements (TEs) are an important type of GSV that comprise over 50% of most eukaryote genomes (Cordaux and Batzer, 2009). TEs are capable of moving in the genome by several mechanisms, including a copy-paste mechanism (Kazazian, 2004). Although many TEs are fixed in the genome of a species—that is, all individuals will have the TE at a specific location in the genome—others are present in some individuals and absent in others, which results in polymorphism at the site of the TE insertion.

Because TEs have important phenotypic consequences on the host genome (Kazazian, 2004; Kapusta et al., 2013; Fedoroff, 2012; Bourque, 2009; Böhne et al., 2008; Kokošar and Kordiš, 2013; O’Donnell and Burns, 2010), it is important to have robust methods to determine the location of a specific element in genomes. These data can be obtained by molecular approaches that amplify the region spanning the end of the TE and the adjacent genome region of the host; a product is obtained only if the TE is present. Multiple methods have been developed to detect different classes of TEs in the genomes of individuals via high throughput sequencing (O’Donnell and Burns, 2010; Iskow et al., 2010), allowing investigators to identify the location of all TEs of a specific type in an individual genome.

Bao et al. (2014) reported recently on a method to detect an endogenous retrovirus (ERV), which is a type of TE derived from an infectious retrovirus, in the genome of the Cervid mule deer (*Odocoileus hemionus*), a species that lacks a reference genome. Each Cervid endogenous retrovirus (CrERV) is present at a unique position in the genome (Elleder et al., 2012; Wittekindt et al., 2010). Animals that share a CrERV must be related because ERVs are inherited along family lineages like any host gene. Thus, animals with a similar profile of CrERV integration sites in their genome have the potential to display similar phenotypic effects of CrERV compared to animals without CrERVs at those locations. In order to investigate the consequences of CrERV integration on the mule deer host, Bao et al. (2014) developed a de novo clustering approach that groups all CrERVs that occupy the same genomic region from different animals. Each cluster of sequences may be represented by a single consensus sequence that in turn represents the site in the host genome where the virus has integrated. The resulting data are an $m \times n$ matrix X , where the (i, j) element X_{ij} gives the count of sequences (the read count) from animal j that are assigned to CrERV-host

junction i , which will henceforth be referred to as virus i . Here, m and n are the total numbers of viruses and animals, respectively.

These read count data contain information about whether an individual carries specific integration sites. However, read counts may contain both false positives and false negatives: A small number of sequences may be attributed to an animal not carrying a particular virus due to either measurement errors in the high-throughput methods or mis-assignment in the clustering process; and no sequences may be captured for an animal actually carrying a particular virus when there are insufficient sequences.

It is therefore challenging to determine the true status of virus i in animal j based on low read counts. One approach is to set a threshold, and assume that a virus is carried by an animal whenever the corresponding read count is above the threshold. This ad-hoc practice has serious drawbacks, as discussed in [Bao et al. \(2014\)](#); essentially, it ignores inherent heterogeneity in the process of generating the counts. Although [Bao et al. \(2014\)](#) move beyond the naive thresholding approach by proposing a mixture model, the mixture used in that article of a Poisson component and a truncated geometric component has several drawbacks. The present article presents a much-improved mixture model, describing the biological and statistical reasons for the modeling choices we make and then discussing the results of fitting this model to the data.

2. A mixture model approach. Count data are sometimes modeled using a Poisson distribution or, if more flexibility is required, a negative binomial distribution. When in addition some of the counts are zeros created by a separate random mechanism, we may introduce a point mass at zero; the resulting “zero-inflated” count models are in fact simplistic mixture models. For our data, zero-inflation is not biologically appropriate because even nonzero counts X_{ij} may occur when virus i is absent from animal j . Instead, we must account for counts in both the “true negative” and “true positive” cases while respecting model parsimony as well as biological realities of the sequencing processes used to obtain the data. This section explains our modeling choices from both a statistical and biological standpoint. In particular, we explain why we have avoided the mixture of Poisson and truncated geometric distributions originally used by [Bao et al. \(2014\)](#).

2.1. The Mixture Model. Let us first consider the situation in which animal j carries virus i , which we might call the “true positive” case. The first model that springs to mind for count data is something based on the Poisson distribution. However, we have found strong evidence of over-dispersion—that is, evidence that the standard deviation of these true positive counts is

larger than the square root of their mean—even when we use a model with a large number of parameters to account for the heterogeneities across animals and viruses. This over-dispersion is depicted in Figure 1, which compares the best-fitting Poisson and negative binomial models in terms of their Pearson residuals, which are the observed counts minus the estimated counts divided by the square roots of the estimated variances. On the other hand, the negative binomial family appears adequate for this modeling task. Figure 1 was created simplistically, by discarding all read counts smaller than 10, both because counts greater than 9 are most likely to be true positives and because it is primarily the large counts that cause the over-dispersion. In the Poisson figures, the estimated counts are based on maximum likelihood estimates in a model assuming that the X_{ij} greater than 9 are distributed independently as $\text{Poisson}(a_i b_j)$ for parameters a_1, \dots, a_m and b_1, \dots, b_n . In the negative binomial figures, the assumption is that the X_{ij} are distributed independently as negative binomial random variables with parameters r_j and α_i , where $1 \leq i \leq m$ and $1 \leq j \leq n$, so that

$$(2.1) \quad P(X_{ij} = x) = f_{ij}(x; r, \alpha) \stackrel{\text{def}}{=} \binom{x + r_j - 1}{x} \alpha_i^{r_j} (1 - \alpha_i)^x$$

for $x = 0, 1, 2, \dots$

Based on Figure 1, the data clearly suggest discarding the Poisson model in favor of the negative binomial model for the true positive mixture component of the model. Interestingly, this choice is not merely in favor of the model with more parameters, as is often the case when a negative binomial distribution fits better than a Poisson distribution; here, each model has the same number of parameters. In equation (2.1), we interpret α_i as a virus-specific parameter where $1 - \alpha_i$ approximates the enrichment of virus i , and r_j as an animal-specific parameter. The mean and variance of the negative binomial distribution of equation (2.1) are $r_j(1 - \alpha_i)/\alpha_i$ and $r_j(1 - \alpha_i)/\alpha_i^2$, respectively. In particular, the mean and variance are both directly proportional to the animal-specific r_j parameter and they are decreasing functions of the virus-specific α_i parameter.

On the other hand, in the “true negative” case where animal j does not carry virus i , in principle we may choose an entirely different class of distributions. We reject the Poisson immediately because we need a distribution with a variance substantially larger than its mean. The geometric distribution is an interesting potential alternative and has the advantage of simplicity since, like the Poisson, it only requires a single parameter. However, we reject the geometric for a different reason: The geometric mass function decays slower for large values than that of a negative binomial, even if the

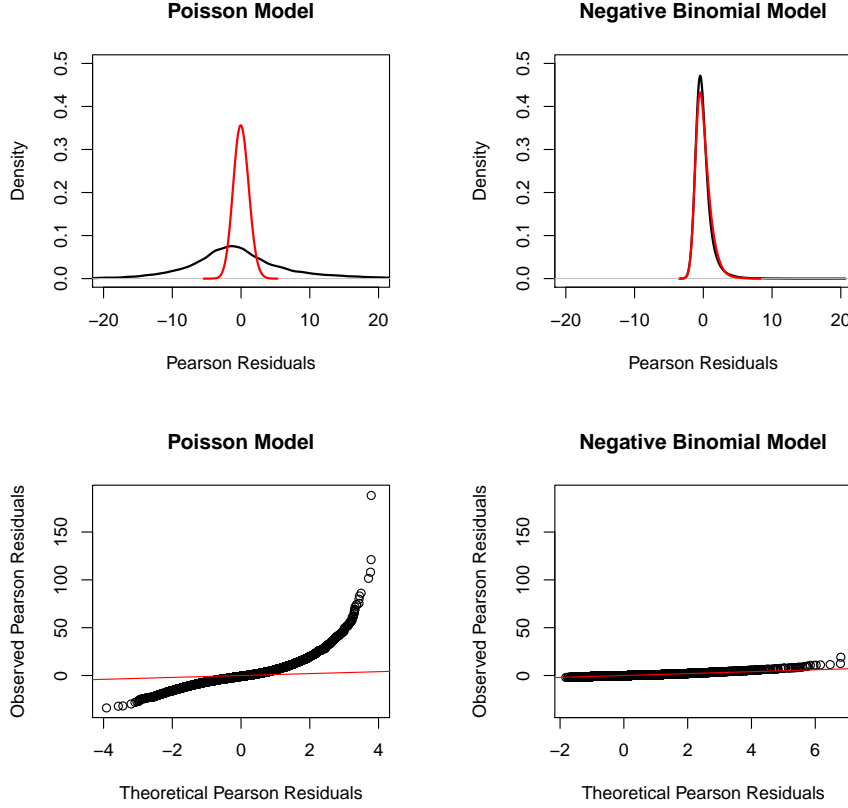


FIG 1. The top two panels show the kernel density estimates of the Pearson residuals (in black) and the theoretical distributions (in red) for the Poisson and negative binomial models. The lower panels are Q-Q plots of the observed vs. theoretical Pearson residuals.

mean of the former is smaller than the mean of the latter, as illustrated in Figure 2. Thus, outlying large counts could be classified as coming from the true negative component, which is nonsensical scientifically. Bao et al. (2014) sidestepped this issue by truncating the geometric distribution of true negative counts. However, we wish to avoid the problematic question of how to choose a truncation point.

Due to the considerations above, we reject both the Poisson and geometric models for the true negative counts in favor of a more flexible negative binomial model and posit that whenever animal j does not carry virus i , the

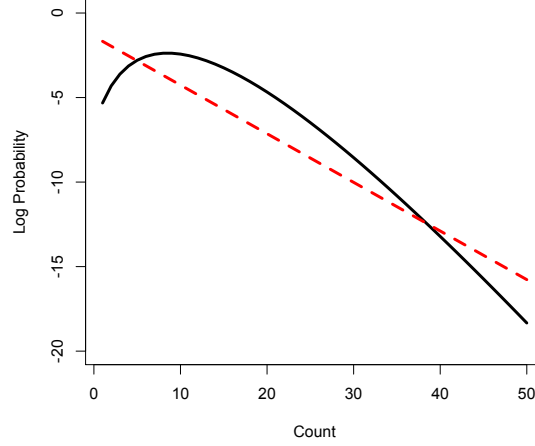


FIG 2. The two mass functions whose logarithms are shown here are a negative binomial with parameters $r = 10$ and $\alpha = 1/2$ (solid, in black) and a geometric with parameter $p = 1/4$ (dashed, in red). The means of the distributions are 10 and 3, respectively, even though the geometric has a larger mass function for large values.

mass function for the count X_{ij} is given by

$$(2.2) \quad g_{ij}(x; r, p) \stackrel{\text{def}}{=} \binom{x + r_j - 1}{x} p_{k(j)}^{r_j} (1 - p_{k(j)})^x, \quad x = 0, 1, 2, \dots$$

where $k(j)$ is the experiment in which animal j was sequenced, r_j is the same animal-specific parameter as in Equation (2.1), and the expected false-positive count for the $k(j)$ experiment is a decreasing function of $p_{k(j)}$: As explained earlier, this expected count is $r_j(1 - p_{k(j)})/p_{k(j)}$.

Both negative binomial distributions, in Equation (2.1) and Equation (2.2), can be interpreted as a sum of r_j independent geometric distributions. This is a deliberate modeling choice that reflects the fact that the quality and quantity of each animal's sample may vary, and this variation will affect counts from both the true positive class and the true negative class in the same way.

The virus-specific effect is most relevant in the true positive case, as reflected by the fact that we allow true positives to depend on the parameter α_i where i denotes the virus number. In the true negative case, counts may be considered to be “background noise” and therefore likely to depend on the particular experiment but not the virus in question; for this reason, we

allow the count distribution for the true negative class to depend on $p_{k(j)}$, where $k(j)$ denotes the experiment number of animal j .

We occasionally obtain distinct sets of counts from the same animal when samples from the same animal are run in different experiments. In such a case, our model treats these counts as though they come from different animals, conditional on the mixture component from which the counts X_{ij} come. That is, each set of counts receives its own index j , so the r_j parameters may be different. This is important since different sets of counts come from distinct experiments, and these often have dramatically different count profiles. In fact, allowing for this flexibility, which is enhanced by indexing the true negative distributions by $p_{k(j)}$, means that our model can easily accommodate new data as they are created in separate sequencing runs or on separate sequencing platforms. This is scientifically important, since our data are continually updated as new animals are sequenced; sequencing technology advances rapidly, and it is not always feasible nor cost-effective to rerun previously sequenced animals using newer technology. Thus, our method allows for seamless data integration by preventing us from having to discard useful data simply because technology changes or our set of sampled animals expands.

On the other hand, it is important that our model can account for cases in which multiple sets of counts come from the same animal in our dataset. This is done by placing appropriate constraints on the mixing probabilities π_{ij} , where π_{ij} represents the a priori probability that animal j carries virus i . Thus, we introduce the constraint $\pi_{ij} = \pi_{ij'}$ for any $j \neq j'$ for which j and j' index the sets of counts from two different runs on the same animal. Once we introduce the π_{ij} probabilities, the full likelihood of our mixture model becomes

(2.3)

$$L(\pi, r, \alpha, p) = \prod_{i=1}^m \prod_{j \in U} \left[\pi_{ij} \prod_{j' \in S_j} f_{ij'}(x; r, \alpha) + (1 - \pi_{ij}) \prod_{j' \in S_j} g_{ij'}(x; r, p) \right],$$

subject to the constraints explained above, where $S_j = \{j' : j \text{ and } j' \text{ are the same animal}\}$ and U is any set containing exactly one element from each S_j ; that is, U is a set of indices for the unique animals. We experimented with three simple parameterizations of the π_{ij} parameters: (1) $\pi_{ij} = \pi$ for all i and j ; (2) $\pi_{ij} = \pi_i$ for all j ; and (3) $\pi_{ij} = \pi_j$ for all i . In Section 3.1, we find that option (2) attains the best Bayesian Information Criterion (BIC) score.

One important implication of the fact that Equations (2.1) and (2.2) represent the true positive and true negative components, respectively, is that biologically we must require that $E(X_{ij})$ is greater in Equation (2.1)

than in Equation (2.2). These two means are given by $r_j(1-\alpha_i)/\alpha_i$ and $r_j(1-p_{k(j)})/p_{k(j)}$, respectively. Thus, since the r_j parameter is common to the two mass functions, the desired inequality may be guaranteed by enforcing the constraints $\alpha_i < p_{k(j)}$ for all i and k during the estimation procedure. To adopt such a strategy in practice would require some potentially ad-hoc choices since even the situation $\alpha_i = p_{k(j)}$ should be avoided. Fortunately, however, we find that our unconstrained point estimates already satisfy $\alpha_i < p_{k(j)}$ when we fit our model to real data, providing support for the modeling choices we have made. We elaborate further on this fact in Section 3.1.

2.2. Parameter Estimation. Estimation of the model parameters is accomplished using maximum likelihood via a straightforward Expectation-Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993). Essentially, an ECM algorithm is merely an EM algorithm in which only one subset of the parameters is updated at each iteration or sub-iteration. In the E-step, given the iteration- t parameter values $\pi_{ij}^{(t)}$, $r_j^{(t)}$, $\alpha_i^{(t)}$, and $p_{k(j)}^{(t)}$, we calculate the probability that animal i carries virus j :

$$(2.4) \quad Z_{ij}^{(t)} = \frac{\pi_{ij}^{(t)} \prod_{j' \in S_j} f_{ij'}^{(t)}(x)}{\pi_{ij}^{(t)} \prod_{j' \in S_j} f_{ij'}^{(t)}(x) + (1 - \pi_{ij}^{(t)}) \prod_{j' \in S_j} g_{ij'}^{(t)}(x)}.$$

In the M-step, we update the parameters in four distinct subsets, in each case holding the other parameters fixed at their most up-to-date values. To wit, we first consider the α parameters. We find that the log-likelihood involving α_i is

$$\sum_{i,j} Z_{ij}^{(t)} \left[\log \left(\frac{x_{ij} + r_j^{(t)}}{x_{ij}} - 1 \right) + r_j^{(t)} \log(\alpha_i) + x_{ij} \log(1 - \alpha_i) \right],$$

which is maximized at

$$\alpha_i^{(t+1)} = \frac{\sum_j Z_{ij}^{(t)} r_j^{(t)}}{\sum_j Z_{ij}^{(t)} (x_{ij} + r_j^{(t)})}.$$

The estimate of α_i will be unstable if Z_{ij} is close to zero for all j , so in practice, we let

$$\alpha_i^{(t+1)} = \frac{\sum_j Z_{ij}^{(t)} r_j^{(t)} + 0.05}{\sum_j Z_{ij}^{(t)} (x_{ij} + r_j^{(t)}) + 0.1},$$

noting that the ascent property guaranteed by an ECM algorithm relies only on the assurance that the complete-data log likelihood increases its value at

each iteration; if the corrected version of $\alpha^{(t+1)}$ ever fails to produce such an increase, it may be replaced by the exact version.

The log-likelihood that involves $p_{k(j)}$ is maximized at

$$p_k^{(t+1)} = \frac{\sum_{j \in A_k} (1 - Z_{ij}^{(t)}) r_j^{(t)}}{\sum_{j \in A_k} (1 - Z_{ij}^{(t)}) (x_{ij} + r_j^{(t)})},$$

where A_k denotes the set of animals coming from the k th experiment. The log-likelihood that involves r_j is maximized at

$$\begin{aligned} r_j^{(t+1)} = \arg \max_{r_j} & \left\{ \sum_i \log \binom{x_{ij} + r_j - 1}{x_{ij}} \right. \\ & \left. + r_j \sum_i \left[Z_{ij}^{(t)} \log \alpha_i^{(t+1)} + (1 - Z_{ij}^{(t)}) \log p_{k(j)}^{(t+1)} \right] \right\}, \end{aligned}$$

which will be solved numerically. Finally, there are several different update formulas for the π_{ij} parameters, depending on which of the three models we are using. We have

$$\pi^{(t+1)} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n Z_{ij}^{(t)}, \quad \pi_i^{(t+1)} = \frac{1}{n} \sum_{j=1}^n Z_{ij}^{(t)}, \quad \text{or} \quad \pi_j^{(t+1)} = \frac{1}{m} \sum_{i=1}^m Z_{ij}^{(t)},$$

depending on the whether we select model (1), (2), or (3), respectively.

We initialize the ECM algorithm at $Z_{ij}^{(0)} = \max(1, x_{ij}/10)$ and $r_j^{(0)} = 100$, and estimate parameters by iterating between the M-step and the E-Step described above.

We stop iterating when the sum of the absolute changes of all $Z_{ij}^{(t)}$ is less than 0.01; these values at convergence will be denoted by \hat{Z}_{ij} ; they represent the probabilities, conditional on the observed data, that animal j has virus i when then the parameter values are taken to be the maximum likelihood estimates. The $m \times n$ matrix of all such probabilities will be denoted \hat{Z} .

Because EM-based algorithms can be sensitive to starting parameter values, we also explore different starting values. Letting $Z_{ij}^{(0)} = \max(1, x_{ij}/c)$ where $c = 2, 3, \dots, 20$, and letting $r_j^{(0)}$ vary from 5 to 500, we find that all these combinations of starting values converge to essentially the same solution.

After the algorithm has converged, the entries of the matrix \hat{Z} may be used as estimates of the probabilistic assignment of viruses to animals, which may in turn lead to insights into the relationships among animals. We revisit this topic in Section 3.3.

3. Results. The 1722×77 matrix X containing the read count data is provided in the supplementary materials. This dataset is an abridged version of the original, excluding any viruses that do not have at least two animals containing five or more sequences. Of all the read counts in the table, 82.6% are zero and another 6.3% are between one and ten, inclusive. The mean of all non-zero counts is 98.6. A second data file, specifying the experiment in which each set of read counts was obtained along with the geographic location (longitude and latitude) where the animal was originally sampled, is also available in the supplementary materials.

3.1. Mixture model parameter estimates. There are three models for the π_{ij} parameters discussed in Section 2.1, and we use the Bayesian information criterion (BIC) to select from among them. Although the original formulation of BIC by Schwarz et al. (1978) is for exponential family models rather than the mixture models we discuss, we only have three models to compare and other model selection criteria such as AIC arrive at the same conclusion since the three models give very different BIC scores, as depicted in Table 1. Model (2) is selected as the best model in both cases, which implies the

Model	Number of model parameters	Treatment of replicates	
		Independent samples	Identical animals
(1) $\pi_{ij} = \pi$	1803	363,673	341,905
(2) $\pi_{ij} = \pi_i$	3524	352,533	336,469
(3) $\pi_{ij} = \pi_j$	1879	361,936	341,906

TABLE 1

Bayesian information criterion (BIC) scores, given by -2 times the maximized log-likelihood plus $mn \log d$, where d is the number of model parameters.

prevalence rates of viruses are heterogeneous. Therefore the following analysis focuses on the $\pi_{ij} = \pi_i$ setting. Since our primary interest is in the matrix \hat{Z} , which is a large matrix, we ignore the calculation of standard errors in this setting. In principle it is possible to exploit asymptotic theory and calculate precision estimates for the r , α , p , and π parameters using a Hessian matrix of the log-likelihood function, yet the resulting 3524×3524 matrix contains many more entries than there are data points and this strategy is therefore unlikely to be productive.

The estimates of the α_i parameters, one for each of the 1722 viruses, range from 1.77×10^{-4} to 0.503, and the estimates of the $p_{k(j)}$ parameters are 0.979, 0.963, and 0.981. We find therefore that $\alpha_i < p_{k(j)}$ in each case, which is biologically necessary as we pointed out in Section 2.1, even though we do not enforce this constraint in the optimization algorithm. This result guarantees that the expected read counts for the “true positive” case are

always larger than those for the “true negative” case. Figure 3 depicts some characteristics of the α_i and r_j estimates in the model treating the replicates as independent samples. The corresponding results for the model treating the replicates as repeats is similar graphically, so we omit it here.

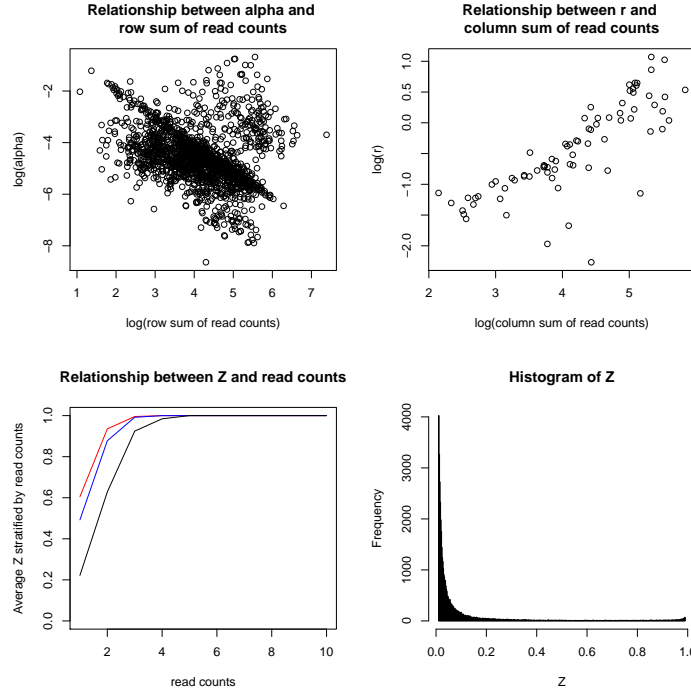


FIG 3. Top row: Row (left) and column (right) means of the X matrix, with zeros omitted, plotted against the corresponding log-parameter estimates, which are the $\log \hat{\alpha}_i$ on the left and the $\log \hat{r}_j$ on the right. Bottom left: Mean \hat{Z}_{ij} values as a function of nonzero values of corresponding X_{ij} values. The three lines indicate stratification by experiment number. Bottom right: Histogram of the \hat{Z} , for $0.01 < \hat{Z} < 0.99$. 52.4% of these probabilities are less than 0.01; and 14.6% of these probabilities are greater than 0.99.

It is worth underscoring several aspects of the \hat{Z}_{ij} estimates. First, the estimated \hat{Z}_{ij} values are not a monotone function of the read counts X_{ij} , which demonstrates that the mixture model approach captures subtle individual heterogeneities among viruses and animals that a simplistic threshold cannot. This fact illustrates that the mixture model is valuable despite the fact that more than 2/3 of all \hat{Z}_{ij} values are close to zero or one (i.e., larger than 0.99 or smaller than 0.01), some values near zero correspond to counts at least as large as those corresponding to some \hat{Z}_{ij} values near one. Finally, the bottom left plot of Figure 3 suggests that for any possible choice of a

threshold on raw sequence counts, even an experiment-specific threshold, will produce either many false negatives (if the threshold is chosen too high) or false positives (if it is chosen too low).

3.2. Validation of mixture model via replicated individuals. One way to get a partial validation of the mixture model in relation to a simplistic count-thresholding method is to consider how well these methods match replicated animals when the replication is ignored during the fitting process. If we let \hat{Y}_{ij} denote a “hard,” or binary, classification of the estimated status of virus i in animal j , then the proportion of consistent \hat{Y}_{ij} across replicates serves as a measure of the accuracy of estimated virus status. There are 11 animals with replicated read counts on 1,722 viruses, allowing us to compare estimated virus status for $1,722 \times 11$ cases. As we change the read count threshold from 1 to 10, 160,000 cases always provide consistent virus statuses, whereas 251 cases always provide inconsistent virus statuses. We are most interested in the remaining 2691 cases where the estimated virus status is sensitive to the choice of threshold, and we summarize the proportion of consistent virus status across replicates for those cases in Figure 4. In the figure, we acknowledge the arbitrary choice of a cutoff value c for defining $\hat{Y}_{ij} = I\{\hat{Z}_{ij} > c\}$ by displaying the results for a range of c values from 0 to 1. When $c = 0$ or $c = 1$, we are essentially ignoring \hat{Z} entirely, so the 100% agreement is uninteresting; yet as the figure shows, every choice of the count threshold is outperformed by the c that yields the same overall proportion of viruses in \hat{Y} . In other words, even if we use the mixture model to perform a hard partition of the read counts into “present virus” and “absent virus” groups, we still outperform the corresponding partition based on a read count threshold.

3.3. Summarizing animal relationships. There are many potential methods to analyze the probabilistic assignment of viruses—or, more generally, TEs and alleles—to animals represented by the \hat{Z} matrix derived from our mixture model and estimation procedure. Broadly speaking, a suite of population genetics tools exists to utilize allele frequency data to estimate population parameters. Several of these methods accommodate probabilistic assignments as well. As an example, Bao et al. (2014) demonstrate a hierarchical clustering method that uses such probabilistic assignments.

Here, we illustrate one type of analysis based on the information provided by the \hat{Z} matrix to estimate how variation in CrERV integration sites is distributed among the animals from the four sampled populations. By considering each column of this matrix as a point in m -dimensional space, we may perform principal components analysis (PCA) and visualize the first

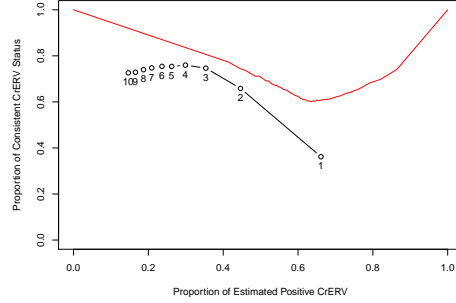


FIG 4. The black line shows the proportion, for count thresholds from 1 to 10 (where a read count at least as large as the threshold is considered a positive virus assignment), of consistent virus assignments to replicated animals as a function of overall proportion of viruses present among the 2691 cases where virus status is sensitive to choice of threshold. The red line shows similar values using the mixture model \hat{Z} matrix and cutoff probabilities ranging from $c = 0$ to $c = 1$. The difference between the two curves ranges from 8.4% to 24.6%.

two principal components. In Figure 5, we see a depiction of the result after the first two PC scores are rotated and scaled so as to make their two-dimensional locations comparable with the geographic locations where the animals were found.

The deer depicted separately from the others in the lower left quadrant of Figure 5 are the blacktail deer subspecies of mule deer that emerged about 20,000 years ago. The close association of Oregon and Montana mule deer to each other and the more distant relationship of Wyoming animals is an unexpected finding, given that previous studies have reported low population subdivision in mule deer.

4. Discussion. The goal of our research was to determine which individuals share a genomic feature, in this case a newly described endogenous retrovirus. Genomic structural variants such as endogenous retroviruses are identified by the position they occupy in the genome. The data used to determine the presence of a variant are often based on the number of reads assigned to the variant. Read count data are heavily skewed toward small numbers, creating uncertainty in the presence/absence status of any particular element. Our article demonstrates the utility of using a mixture model to assign a probability that a virus is present in a given animal. Because these viruses are inherited like any host gene, animals that share more viruses are more closely related. Our results show that animals from Wyoming can be

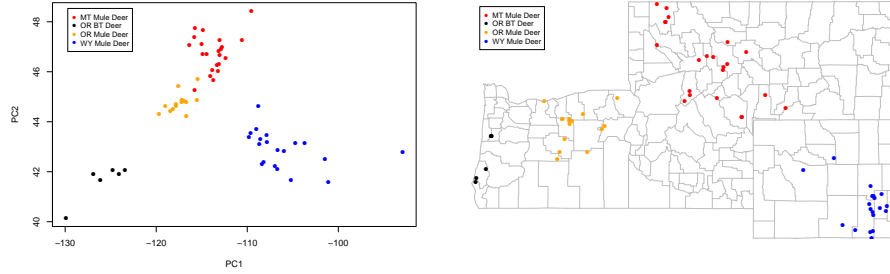


FIG 5. *First two principal component scores (left) and geographic locations (right) of the deer. In the legend, BT stands for blacktail and MT, OR, and WY stand for Montana, Oregon, and Wyoming.*

distinguished from those from the adjacent state of Montana based on the profile of shared virus integration sites. This is a surprising finding because mule deer are migratory animals and can move between these two geographic locations. In fact, based on these analyses, the Montana mule deer appear more closely related to those in Oregon. Studies using traditional approaches report that mule deer have little population structure throughout this region (Latch et al., 2014).

While we demonstrate the utility of using a mixture model for read count data for an endogenous retrovirus, our methodology is applicable to any data aimed to determine the presence or absence of a polymorphic element—for instance, a different class of mobile element such as a long interspersed nuclear element, or LINE—at a specified position in the genome (Akagi et al., 2008; Evrony et al., 2012; Burns and Boeke, 2012; Richardson, Morell and Faulkner, 2014). These would be present at different frequencies, unlike genomic elements, which would affect the read count distribution.

The primary statistical contributions of this article are twofold: First, it reinforces and provides additional evidence to support the argument made in Bao et al. (2014) that a two-component mixture model for estimating probabilities of binary outcomes being positive, given observed count data, is more flexible, principled, and accurate than the commonly-used approach of dichotomizing results based on a count threshold. Second, it significantly advances the mixture approach proposed by Bao et al. (2014) by carefully considering statistical and biological features of these data. As one indication that the fitted model gives biologically sensible results, we find that in all cases, the best-fitting parameters imply that $E(X_{ij}|j \text{ contains } i) > E(X_{ij}|j \text{ does not contain } i)$ even though we do not enforce this inequality

using constraints.

Our approach has the additional feature that it allows seamless integration of data from multiple experiments. This is prudent because not all samples included in an analysis are processed at the same time. Biological realities such as different “true negative” characteristics for different experimental runs and samples that are replicated in more than one experiment can be automatically accounted for by the model. As a case in point, the read counts we analyze in this article are a superset of the counts used by Bao et al. (2014).

In our dataset, the counts from multiple experiments all used the same Ion Torrent sequencing platform; yet in principle the model we propose can incorporate data from different platforms as well, which is important because sequencing technology advances rapidly and so techniques such as ours that do not necessitate discarding “old” runs are both scientifically prudent and economical. Indeed, the adoption of our method enables the experimenter to consider designing experiments that include some replicated animals between experiments since this overlap will serve to validate the results. This leads to further questions of how to design such experiments optimally to achieve the best tradeoff of statistical accuracy and experimental cost, which could be considered in future work.

References.

- AKAGI, K., LI, J., STEPHENS, R. M., VOLFOVSKY, N. and SYMER, D. E. (2008). Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Research* **18** 869–880.
- BAO, L., ELLEDER, D., MALHOTRA, R., DEGIORGIO, M., MARAVEGIAS, T., HORVATH, L., CARREL, L., GILLIN, C., HRON, T., FÁBRYOVÁ, H., HUNTER, D. and POSS, M. (2014). Computational and Statistical Analyses of Insertional Polymorphic Endogenous Retroviruses in a Non-Model Organism. *Computation* **2** 221–245.
- BÖHNE, A., BRUNET, F., GALIANA-ARNOUX, D., SCHULTHEIS, C. and VOLFF, J.-N. (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome research* : **16** 203–15.
- BOURQUE, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current opinion in genetics & development* **19** 607–12.
- BURNS, K. H. and BOEKE, J. D. (2012). Human transposon tectonics. *Cell* **149** 740–52.
- CORDAUX, R. and BATZER, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics* **10** 691–703.
- ELLEDER, D., KIM, O., PADHI, A., BANKERT, J. G., SIMEONOV, I., SCHUSTER, S. C., WITTEKINDT, N. E., MOTAMENY, S. and POSS, M. (2012). Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. *Journal of virology* **86** 2787–96.
- EVRONY, G. D., CAI, X., LEE, E., HILLS, L. B., ELHOSARY, P. C., LEHMANN, H. S., PARKER, J. J., ATABAY, K. D., GILMORE, E. C., PODURI, A., PARK, P. J. and WALSH, C. A. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151** 483–496.

- FEDOROFF, N. V. (2012). Transposable Elements , Epigenetics , and Genome Evolution. *Science* **338** 758–67.
- ISKOW, R. C., MCCABE, M. T., MILLS, R. E., TORENE, S., PITTARD, W. S., NEUWALD, A. F., VAN MEIR, E. G., VERTINO, P. M. and DEVINE, S. E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141** 1253–61.
- KAPUSTA, A., KRONENBERG, Z., LYNCH, V. J., ZHUO, X., RAMSAY, L., BOURQUE, G., YANDELL, M. and FESCHOTTE, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS genetics* **9** e1003470.
- KAZAZIAN, H. H. (2004). Mobile elements: drivers of genome evolution. *Science (New York, N.Y.)* **303** 1626–32.
- KOKOŠAR, J. and KORDIŠ, D. (2013). Genesis and Regulatory Wiring of Retroelement-Derived Domesticated Genes: A Phylogenomic Perspective. *Molecular Biology and Evolution* **30** 1015–1031.
- LATCH, E. K., REDING, D. M., HEFFELFINGER, J. R., ALCALÁ-GALVÁN, C. H. and RHODES, O. E. (2014). Range-wide analysis of genetic structure in a widespread, highly mobile species (*Odocoileus hemionus*) reveals the importance of historical biogeography. *Molecular ecology* **23** 3171–3190.
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* **80** 267–278.
- O'DONNELL, K. A. and BURNS, K. H. (2010). Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mobile DNA* **1** 21.
- RICHARDSON, S. R., MORELL, S. and FAULKNER, G. J. (2014). L1 retrotransposons and somatic mosaicism in the brain. *Annual Review of Genetics* **48** 1–27.
- SCHWARZ, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics* **6** 461–464.
- WITTEKINDT, N. E., PADHI, A., SCHUSTER, S. C., QI, J., ZHAO, F., TOMSHO, L. P., KASSON, L. R., PACKARD, M., CROSS, P. and POSS, M. (2010). Nodeomics: pathogen detection in vertebrate lymph nodes using meta-transcriptomics. *PloS one* **5** e13432.